LA-UR-18-22949
Approved for public release; distribution is unlimited.

Title: High Speed Networks 101

Author(s): Coulter, Susan K.

Intended for: Presentation at recruitment fairs and/or other student-focused events.

Issued: 2018-04-06

HSN 101

Susan Coulter

HPC-DES / Infrastructure Team

2018-03-05

skc@lanl.gov

History

- Need for high bandwidth, low latency network grows in the late 1990's
- Two designs by two different groups of folks

  (IBM, HP, Compaq vs Intel, MS, Sun)
- 1999 these folks got together and developed InfiniBand (IB)
- The InfiniBand Trade Association (IBTA) was formed
- Standards and specifications so all hardware and software worked together
- Early offering called IBGold – installed on LANL's coyote cluster

History – continued

- OpenFabrics Alliance (OFA) formed in 2004 – national laboratories and vendors
- Name changed from IBGold to OFED (Open Fabrics Enterprise Distribution)
- OFED v1.1 released in 2005/2006
- Several versions of OFED were released
- Interaction with Linux kernel folks was messy
- Significant effort put into getting OpenFabrics Software into the distros
- RedHat (and CentOS) began bundling the code

History – recent

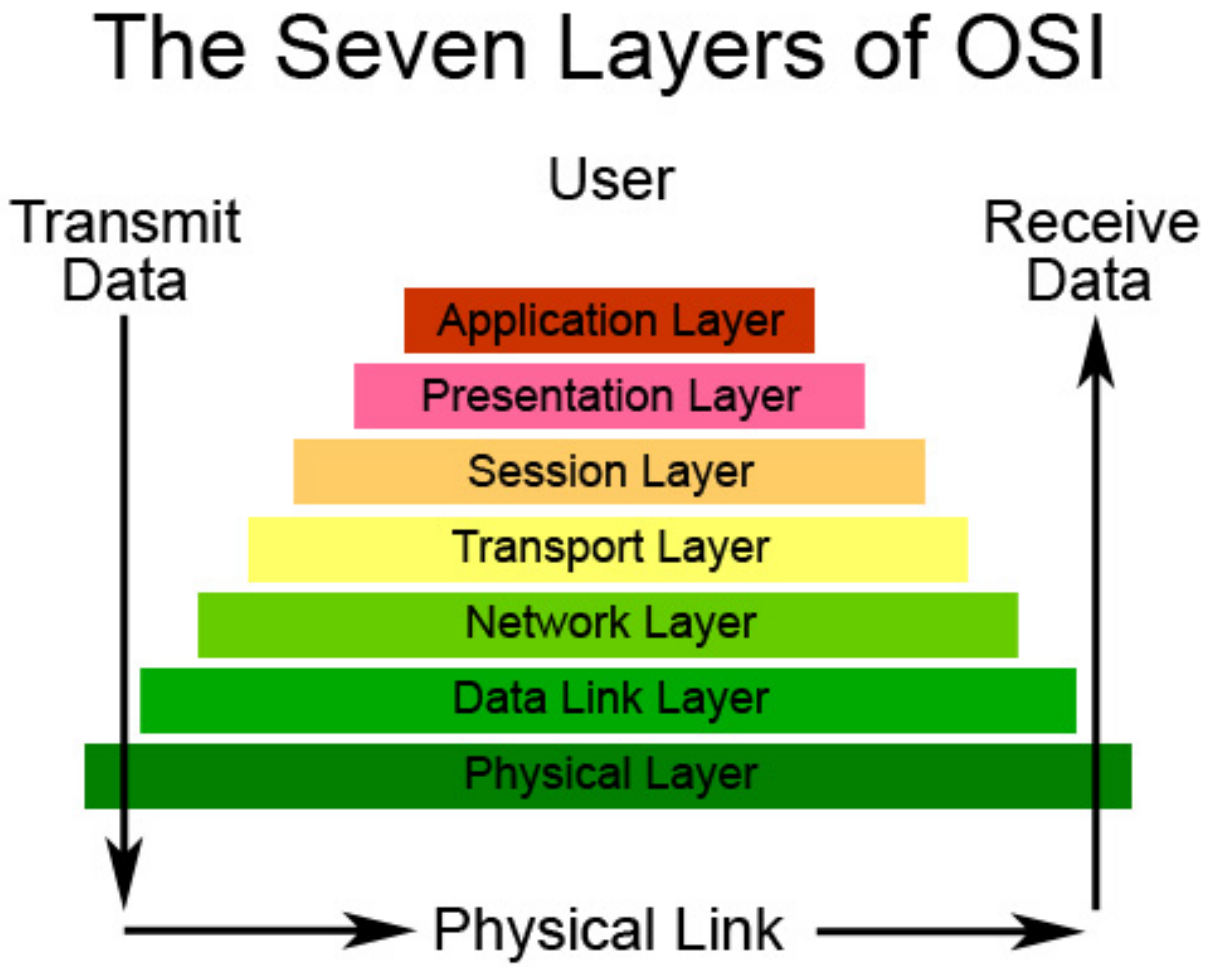
- OpenFabrics Software was accepted and expected in the distros
- RedHat, CentOS, SUSE, Debian, Ubuntu
- Linux kernel community became more accepting and welcoming
- Code considered “fully open source” and moved to GitHub from OFA website
- OFED still exists, but is an “add on” package not necessary in most instances
- Mellanox off-load versus QLogic on-load (Verbs versus PSM(2))
- Intel released Omni-Path (OPA) – after buying QLogic intellectual property

How does it do that ?

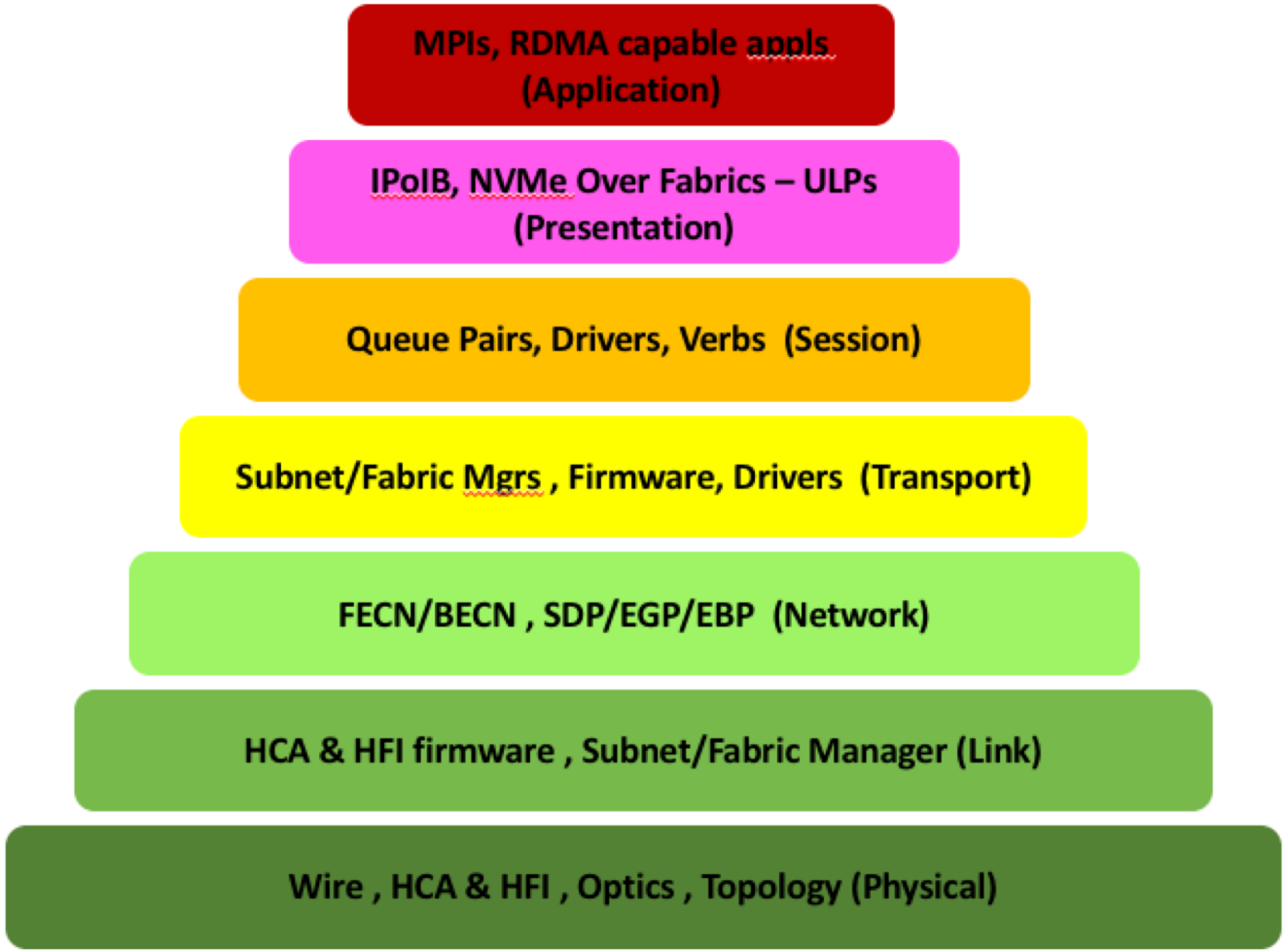
- Layer 2 / switched point-to-point network (link layer)
- CPU offload – in the Verbs case
- Remote Direct Memory Access – RDMA
- Memory regions assigned/pinned to HSN interface cards
- Low processing overhead (no tcp stack)
- Well engineered hardware – once it is mature (IB versus OPA)

- ! IB IS NOT ETHERNET !

Ethernet OSI 7 layer model



IB/OPA Equivalent of 7 layer model



Physical Layer Specifics

InfiniBand

- Can have active or passive cables
- Data is striped across 4 links – 4x links (or 12x)
- Each link runs at the speed of the hardware in use (QDR, FDR, EDR, etc)
- Originally, 8/10 bit encoding, 8 bits of data + 2 bits control = 10
- Newer technology (FDR, EDR) uses 64/66 … ~97% data versus 80% data

Omni-Path

- Can have active or passive cables
- Data is striped across 4 links – 4x links (or 12x)
- Each link runs at the speed of the hardware in use (100Gb, etc)

Link Layer Specifics

- IB/OPA is a layer 2 system – a fabric
- Link flow control and CRC
- No packet fragmentation – packet frames contain SDP and EGP or EBP
- Max '*base*' IB MTU is 4k – HCA can negotiate
- Max '*base*' OPA MTU is 8k
- Ports on a switch do not need configuration
- Activating any port on the fabric sends SNMP trap
- Every port has a path to every port
- Subnet/Fabric Manager receives SNMP traps and creates routes

Network Layer Specifics - Congestion Control

- Some of our clusters have experienced congestion issues
- Data is put on the wire when enough ‘credits’ exist for that port
- Bad routing algorithms can cause ‘credit loops’ - deadlocks
- End to end congestion control newly available
- Called “lossless”, but congested fabrics do drop packets
- XmtDiscards

Transport/Session Layer Specifics

- No TCP stack
- Think in messages, not packets
  - TCP rebuilds fragmented packets
  - IB/OPA handles fragmented messages
  - Some applications can handle out-of-order messages, some cannot
- IB Queue Pairs for manage the communications
- QP State Machine off-loaded to the HCA
- OPA uses “contexts” and uses one of the on board processors

Transport Layer continued
Subnet/Fabric Manager – the original SDN

- Switches are really dumb – ‘static’ routing table
- Routes are created by a Subnet/Fabric Manager
- Given a topology and algorithm, manager creates Linear Forwarding Table
- Sends that table to every switch in the fabric
- Sweeps the fabric at a configurable granularity of time
- Link state changes can prompt modification of routes
- Uses special management packets
- IB mad packets transferred on special Virtual Lane, VL15 – no flow control which avoids deadlocks

Presentation Layer Specifics – IPoIB example

- IP emulation layer using underlying HSN messages
- Used by many software packages, including Lustre and OpenMPI, and external file access
- Connected mode or Datagram mode set fabric wide
  - Connected has larger MTU and better validation
  - Datagram has smaller MTU, lacks ACK and other checking
- arp used in a similar manner – requires 4 messages
    - arp broadcast
    - Path Record (PR) request
    - PR response
    - arp unicast response
- Appears on *ifconfig* command results
- Unicast and Multicast

Terms & Concepts

- Chassis vs Switch terminology Physical box is called a chassis, contains multiple switches
- Spine Cards (or Spine Switch/Fabric Board)
- Line Cards (or Leaf Switch)
- QDR, FDR, EDR - 36 ports per switch
- OPA1 – 48 ports per switch
- HDR will have 40 ports of 200G or 80 ports of 100G
- Line card switches – half go to the backplane/spine and half are external
- External ports connect to hosts or other switches

More Terms & Concepts

- GUID / LID – IB/OPA world revolves around these
- GUID is 128bit mac address / LID is transient integer – one to one mapping
- SLID / DLID – IB/OPA uses destination routing

InfinBand

- Up to 16 Virtual lanes including VL15 for management control
- Lid Mask Control (lmc) – multiple routes per HCA
- Service Levels / Quality of Service (SL / QOS)
- Subnet Manager (SM)
- SMA/SA – subnet (manager) administrator

  Runs on every node/switch

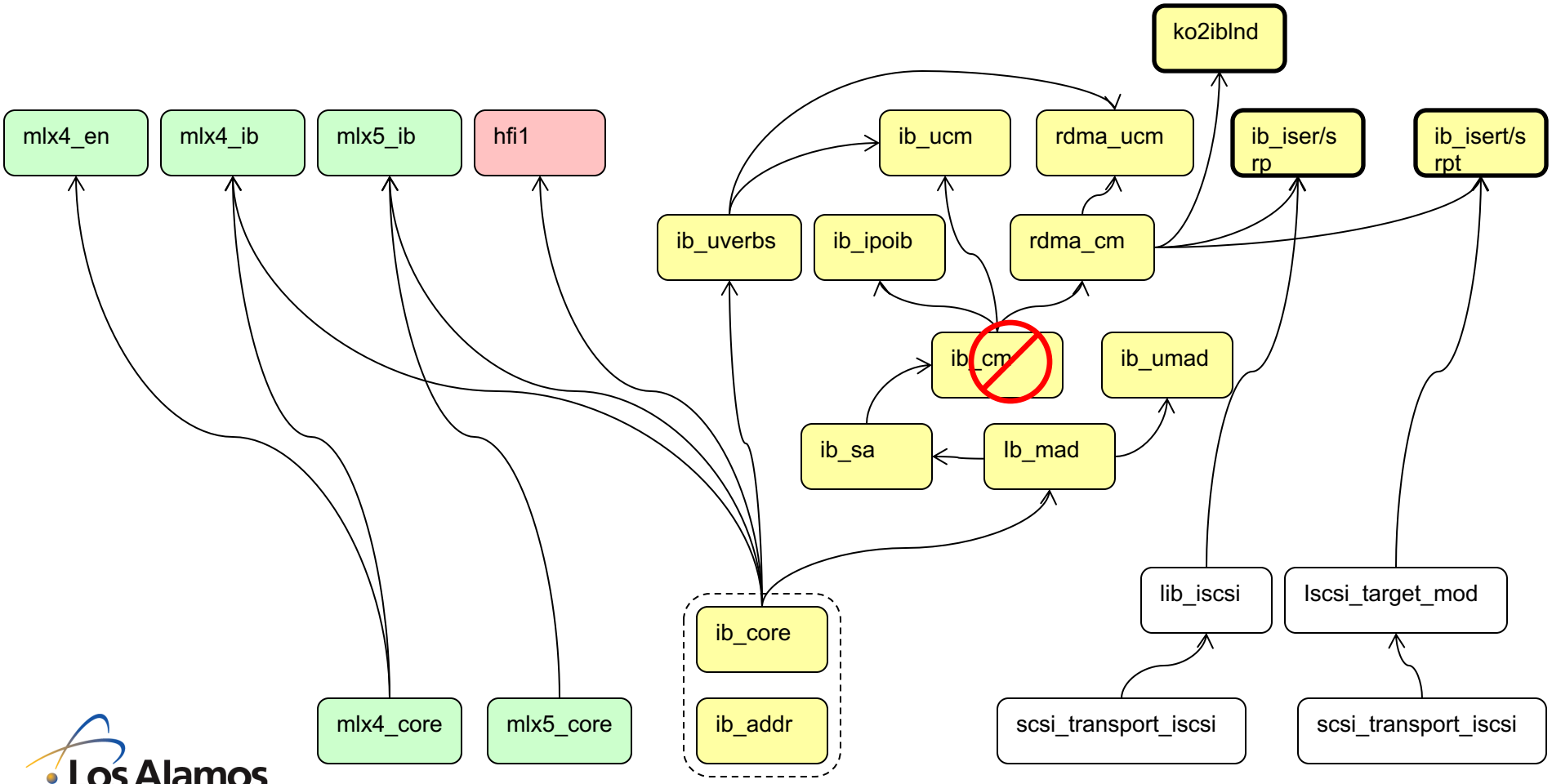
More Terms & Concepts

- GUID / LID – IB/OPA world revolves around these
- GUID is 128bit mac address / LID is transient integer – one to one mapping
- SLID / DLID – IB/OPA uses destination routing

Omni-Path

- Service Levels / Quality of Service (SL / QOS)
- Fabric Manager (FM)
- Congestion control is very fine grained, can modify mid-transmission

Kernel View HSN



High Speed Network 101

End

Thank You